Consumer Income and Expenditure Division

Division du revenu et des depenses des consommateurs

L-0423

Statistics Canada

Consumer Income and Expenditure Division

Family Expenditure Section

Statistique Canada

Division du revenu et des dépenses des consommateurs

Section des dépenses des familles

An Investigation of Variance Estimation in Family Expenditure Surveys

An Empirical Study

L'estimation de la variance dans les enquêtes sur les dépenses des familles

Étude empirique

Published under the authority of the President of the Treasury Board

Statistics Canada should be credited when reproducing or quoting any part of this document

© Minister of Supply and Services Canada

February 1980
8-3301-515

Ottawa

Publication autorisée par le président du Conseil du Trésor

Reproduction au citation autorisées sous réserve d'indication de la source: Statistique Canada

© Ministre des Approvisionnements et Services Canada

Février 1980
8-3301-515

Ottawa

PREFACE

Statistics Canada utilizes for its household surveys nearly exclusively the Labour Force Survey sampling frame. A detailed description of this survey design, which is a multi-stage stratified clustered area probability sample, can be found in Methodology of the Canadian Labour Force Survey, 1976, Catalogue 71-526. Sampling errors from such samples are complex to estimate and are in magnitude usually larger than those from a simple random sample of the same size.

For standard statistics like proportions and means, sampling errors are published (or are available upon request) that have been calculated based on the methodology originally developed by Keyfitz. For this method the total sample is delineated into a number of paired areas based on the sample design; areas are paired on the basis of similarities in their socio-economic characteristics. Differences in statistics calculated for the two components of these paired areas are used to obtain estimates of sampling variability. Such calculations are cumbersome to perform and suffer from inadequate degrees of freedom in the case of more complex statistics such as regression coefficients and Gini coefficients.

This paper explores alternative methods for estimating the variance of proportions, means and regression coefficients. Generally researchers use sample observations in regressions as if they came from a simple random sample. The availability and efficiency of package computer programs providing regression routines, on the one hand, and the availability of non-identifiable microdata, on the other hand, seem to have encouraged the spread of this practice.

This paper demonstrates the differences in variance estimates that result from various methods of calculation that in turn reflect some different assumptions about the underlying sample. The data utilized come from the 1974 Family Expenditure Survey. This survey was slightly more complex in terms of sample design than other household surveys which

PRÉFACE

Statistique Canada utilise pour ses enquêtes sur les ménages presque exclusivement le plan de sondage de l'enquête sur la population active. Une description détaillée de ce plan de sondage, qui est un échantillon probabiliste stratifié à plusieurs degrés d'une région constituée en grappes, se trouve dans Méthodologie de l'enquête sur la population active du Canada, 1976 (n° 71-526 au catalogue). Les erreurs d'échantillonnage sont difficiles à estimer et leur grandeur est en général plus importante que celle des erreurs d'un échantillon aléatoire simple de même taille.

Dans le cas des statistiques ordinaires, telles que les proportions et les moyennes, on publie (ou l'on met à la disposition des utilisateurs) des erreurs d'échantillonnage calculées selon la méthode mise au point à l'origine par Keyfitz. D'après cette méthode, l'échantillon total est découpé en un certain nombre de régions couplées à partir du plan de sondage; les régions sont couplées selon les similitudes de leurs caractéristiques socio-économiques. Les différences entre les statistiques calculées pour les deux composantes de ces régions couplées servent à calculer des estimations de la variabilité de l'échantillonnage. Ces calculs sont laborieux et sont pénalisés par des degrés de liberté inappropriés dans le cas des statistiques plus complexes, telles que des coefficients de régression ou les coefficients de Gini.

L'auteur examine d'autres méthodes d'estimation de la variance des proportions, des moyennes et des coefficients de régression. De façon générale, les chercheurs se servent des observations de l'échantillon dans les régressions comme si elles provenaient d'un échantillon aléatoire simple. Il semble que cette façon d'agir s'explique d'une part par l'existence et l'efficacité des ensembles de programmes informatiques offrant des sous-programmes de régression, et la disponibilité de microdonnées non identifiables, d'autre part.

Ce document fait ressortir les différences des estimations de la variance imputables aux diverses méthodes de calcul qui rendent elles-mêmes compte de différentes hypothèses sur l'échantillon en cause. Les données proviennent de l'enquête sur les dépenses des familles de 1974. Cette enquête a été légèrement plus complexe pour ce qui est du plan de sondage que d'autres enquêtes sur les mé-

usually utilize the labour force sample frame without any further adjustments. The resulting estimates are, however, still illustrative of the differences that the various calculation methods make. Clearly, the subsequent significance tests that researchers would normally perform are affected by these differences and can lead to wrong conclusions.

Although the estimates of sampling variability presented in this paper are of particular interest to users of the 1974 Family Expenditure Survey, it is hoped that the paper will serve a broader purpose. It provides references and guidance to techniques of estimating variances that are appropriate for the sample design utilized by most Statistics Canada household surveys and that are particularly useful in the context of regression analysis. Further, it is obvious from other research that the same techniques (e.g., jackknife repeated replication method) lend themselves to adaptation in case of other complex statistics such as the Gini ratio.

The paper was written by Mr. John Chadwick from the Family Expenditure Survey Section, Consumer Income and Expenditure Division, in consultation with his colleagues. The assistance of the Methodology Division staff in providing comments and suggestions on earlier drafts is also being acknowledged with special thanks to Mr. L. Swain.

nages qui utilisent habituellement tel quel le plan de sondage de la population active. Les estimations obtenues, toutefois, montrent une nouvelle fois la différence qu'entraîne l'emploi de méthodes différentes. Il est clair que les tests d'hypothèse subséquents que les chercheurs feraient normalement s'en trouvent affectés, ce qui peut mener à des conclusions erronées.

Bien que les estimations de la variabilité de l'échantillonnage présentées ici intéressent particulièrement les utilisateurs de l'enquête sur les dépenses des familles de 1974, l'auteur espère atteindre un but plus vaste. Il fournit des références et des conseils en ce qui concerne les techniques d'estimation de la variance qui conviennent pour le plan de sondage retenu dans la plupart des enquêtes sur les ménages de Statistique Canada, et qui sont particulièrement utiles dans le cas de l'analyse de régression. De plus, il apparaît, à la lumière d'autres travaux de recherche, que les mêmes techniques (telles que la méthode de la subdivision répétée dite du "jackknife") se prêtent à des adaptations dans le cas d'autres statistiques complexes, comme le rapport de Gini.

L'auteur est M. John Chadwick, de la Section de l'enquête sur les dépenses des familles de la Division du revenu et des dépenses des consommateurs. Il a écrit ce document en consultation avec ses collègues. Il tient à remercier le personnel de la Division de la méthodologie pour leurs observations et suggestions à propos de versions antérieures, et tout particulièrement M. L. Swain.

TABLE OF CONTENTS

Section	Page
I. Introduction	7
II. Estimation Procedures	11
The Traditional Approach	
Balanced Repeated Replications	
Jackknife Techniques	
Regression Coefficients	
III. The Results	21
Sample Means	
Regression Coefficients	
Summary and Conclusions	
Text Table	
I. Estimates of Standard Error of Mean	22
II. Regression Coefficients and Estimated Standard Errors Dependent Variable: Food Expenditure	24
III. Regression Coefficients and Estimated Standard Errors Dependent Variable: Shelter Expenditure	26
IV. Regression Coefficients and Estimated Standard Errors Dependent Variable: Clothing Expenditure	28
Bibliography	33

TABLE DES MATIÈRES

Section	Page
I. Introduction	7
II. Procédures d'estimation	11
L'approche traditionnelle	
Subdivision répétée équilibrée	
Les techniques du jackknife	
Coefficients de régression	
III. Les résultats	21
Moyennes de l'échantillon	
Coefficients de régression	
Résumé et conclusions	
Tableau explicatif	
I. Estimations de l'erreur-type de la moyenne	22
II. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses en aliments	24
III. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses pour l'hébergement	26
IV. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses en vêtements	28
Bibliographie	33

SECTION I

INTRODUCTION

Over a period of time, numerous users of family expenditure data have raised questions which have caused the problem of variance estimation to come under scrutiny. This is true both with the variances of sample means and of regression coefficients. These cases are important for construction of significance tests and for determination of the reliability of averages and coefficients.

The basic problem with estimation of variances of family expenditure results is that the sample design for the surveys is derived from the Labour Force Survey sampling framework; thus, it is essentially a stratified multistage clustered scheme. Sampling errors from a complex sampling method are usually difficult to determine, and are generally higher than the sampling errors of simple random samples of the same size. Estimates of standard errors have been traditionally obtained and published based on a method which uses the differences in estimates for two components of a number of designated paired areas into which the sample is delineated. It is, however, our concern that some data users have requested special analytical results which use package computer programs to obtain output such as standard deviations, coefficients of variation and standard errors of regression coefficients which, having made no allowance for the sample design, assume a simple random sample.

The accurate estimation of standard errors is essential for the construction of tests of significance. Further, we are interested in documenting the differences in estimates produced by the various procedures proposed in the literature. Changing the estimation procedure amounts to making different assumptions about the sampling technique. Most package computer programs assume a simple random sample from an infinite universe but as more complicated sampling schemes are used, different variance estimation techniques are required. Having documented the differences in estimates, it will be possible

SECTION I

INTRODUCTION

Avec le temps, de nombreux utilisateurs des données sur les dépenses des familles ont posé des questions qui ont amené un examen plus attentif du problème de l'estimation de la variance. C'est vrai à la fois pour les variances des moyennes des échantillons et des coefficients de régression. Cette question est importante pour la construction de tests d'hypothèse et la détermination de la fiabilité des moyennes et des coefficients.

Le problème fondamental de l'estimation des variances des dépenses des familles est que l'échantillon des enquêtes est établi à partir de la base de sondage de l'enquête sur la population active. Il s'agit donc essentiellement d'un échantillon stratifié multiple en grappes. La détermination des erreurs d'échantillonnage d'une méthode aussi complexe est souvent difficile, et les erreurs d'échantillonnage sont en général plus élevées que celles d'échantillons aléatoires simples de même taille. Les estimations des erreurs-types ont été de façon générale obtenues et publiées grâce à une méthode faisant intervenir les différences des estimations de deux composantes d'un certain nombre de régions couplées désignées servant à la définition de l'échantillon. Nous nous préoccupons toutefois du fait que certains utilisateurs de données ont demandé des résultats analytiques spéciaux qui utilisent des ensembles de programmes informatiques afin d'obtenir des données telles que les écarts-types, les coefficients de variation et les erreurs-types de coefficients de régression, programmes qui, ne prenant pas en compte le plan de sondage, supposent un échantillon aléatoire simple.

Il est essentiel de disposer d'estimations précises des erreurs-types afin de pouvoir effectuer des tests d'hypothèse. De plus, nous voulons documenter les différences des estimations obtenues grâce aux différentes méthodes connues. Changer la procédure d'estimation équivaut à faire différentes hypothèses sur la technique d'échantillonnage. La plupart des ensembles de programmes informatiques supposent un échantillon aléatoire tiré d'un univers infini, mais à mesure que l'on utilise des plans de sondage plus complexes, il faut se servir de techniques différentes d'estimation de la variance. Une fois les différences d'estimations documentées, il sera possible de

to provide any technical advice required by the users of our data.

The main purpose of this report is to quantify the differences between the sampling errors assuming simple random sampling and the estimates which take into account the actual design and estimation procedures. To do this, the traditional method suggested above will be compared with the simple random sample method and with the methods of balanced repeated replication and jackknife repeated replication as outlined in Frankel(2).

The methodologies for the estimation procedures are not new. They are discussed and referenced in Section II of this report. This is, however, the first time that the balanced repeated replication and jackknife techniques have been applied to Canadian family expenditure data. The main thrust of the report will be aimed at a comparison of the statistics produced by package regression programs and the statistics produced by calculations which take into account the actual sample design.

The procedures which are discussed will be demonstrated using the 1974 Survey of Family Expenditures. This survey, conducted in early 1975, provided a usable sample of 6,630 spending units which existed for the full year, from 14 major Canadian cities. The survey was conducted by asking selected families to recall and record all their financial transactions of the calendar year 1974. The sample was broken down into two parts. The first was the regular sample comprising 5,389 spending units, and the second a special low income sample of 1,241 respondents obtained by using a screening procedure. The results of the survey were weighted to accord each record its relative influence in the population of the cities and to provide a balance between the regular and low income parts of the survey which accurately reflects the population. Further details of this methodology are found in the Statistics Canada publication "Urban Family Expenditure 1974", Catalogue 62-544(8).

As the procedures have been tested specifically with the 1974 family expenditure data, the results pertain to that survey only. It is anticipated by the author, however, that further work with other data derived from surveys using the Canadian labour force frame will produce similar results. More empirical work is required to answer this question.

fournir tout conseil technique que les utilisateurs de nos données pourront demander.

Le but principal de ce document est de quantifier les différences entre les erreurs d'échantillonnage dans l'hypothèse d'un échantillon aléatoire simple et les estimations qui prennent en compte le plan de sondage réel et les procédures d'estimation. Pour ce faire, nous allons comparer la méthode traditionnelle suggérée ci-dessus à la méthode faisant intervenir l'échantillonnage aléatoire simple et à celles de la subdivision répétée équilibrée et du jackknife décrites dans Frankel(2).

Les méthodologies des procédures d'estimation ne sont pas nouvelles. Elles sont examinées (avec références) dans la deuxième section. Il s'agit toutefois de la première application des méthodes de la subdivision répétée équilibrée et du jackknife aux données sur les dépenses des familles canadiennes. L'auteur s'attachera particulièrement à une comparaison des statistiques produites par des ensembles de programmes de régression et des statistiques obtenues grâce à des calculs faisant intervenir le plan de sondage réel.

Les procédures examinées seront appliquées aux données de l'enquête sur les dépenses des familles de 1974. Cette enquête, effectuée au début de 1975, a donné un échantillon utilisable de 6,630 unités de dépense qui existaient tout au long de l'année, dans 14 grandes villes canadiennes. On a effectué l'enquête en demandant à certaines familles d'enregistrer toutes leurs opérations financières de l'année civile 1974. L'échantillon a été partagé en deux. La première partie regroupe l'échantillon normal de 5,389 unités, et la deuxième, un échantillon spécial de 1,241 répondants à faibles revenus, obtenu par criblage. Les résultats de l'enquête ont été pondérés de façon à attribuer à chaque dossier son importance relative dans la population des villes et de trouver le point d'équilibre entre les parties normale et à faible revenu de l'enquête qui rendent compte correctement de la population. D'autres renseignements sur la méthode utilisée se trouvent dans la publication de Statistique Canada "Dépenses des familles urbaines, 1974" (n° 62-544 au catalogue)(8).

Comme on a testé les procédures uniquement pour les données sur les dépenses des familles de 1974, les résultats ne sont valables que pour cette enquête seulement. L'auteur prévoit, toutefois, que ces calculs sur des données obtenues à partir d'enquêtes utilisant le plan de sondage de la population active canadienne vont donner des résultats semblables. D'autres travaux empiriques sont nécessaires.

This report is arranged as follows. Section II discusses the statistical techniques underlying the presently used standard error estimation scheme, the balanced repeated replication procedure, and the jackknife method. The methods of applying the latter two to regression estimates are also discussed. Section III presents, in tabular form, results of the techniques outlined in Section II. A reader who is familiar with the methodology could turn directly to Section III.

Voici le plan de ce document. La Section II examine les techniques statistiques propres à la méthode actuellement utilisée des estimations de l'erreur-type, la méthode de la subdivision répétée équilibrée et celle du jackknife. On examine également les façons d'utiliser ces deux dernières dans les estimations de régression. La Section III présente sous une forme tabulaire les résultats des méthodes décrites dans la section précédente. Le lecteur bien au fait de la méthodologie peut passer directement à la Section III.

SECTION II

ESTIMATION PROCEDURES

When estimating variances from sample surveys the procedure to follow is determined by the sampling scheme used to collect the data and by the analysis procedures used. The design of the sample will necessitate on occasion the use of approximate formulae and procedures if exact ones are not available or if they prove to be too intractable, in a mathematical sense, to develop. In this section five procedures will be presented, and their uses explained in terms of sample means and regression coefficients. In all cases the finite population correction will be dropped, assuming a small sampling ratio.

If a simple random sample is assumed the estimator

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

can be used to estimate the mean of a variate X where n is the sample size, x_j, $j = 1 \ldots n$ are the observations. A corresponding variance estimator is the well-known expression

$$v(\bar{x}) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

This quantity has well understood, nicely behaved properties. It is an unbiased estimator of the true sample variance σ^2 and tends to a chi square distribution asymptotically as n increases.

Unfortunately this expression is only valid when a simple random sample has been used. In practice, however, many sample designs are stratified. The use of stratification invalidates both of the above expressions. The use of sample weights will remedy this problem. The weights are numbers associated with each observation

SECTION II

PROCÉDURES D'ESTIMATION

Lors de l'estimation des variances des enquêtes par sondage, la procédure à suivre est déterminée par le plan de sondage utilisé pour la collecte des données et par les procédures d'analyse employées. La conception de l'échantillon nécessitera parfois l'utilisation de formules et de procédures approximatives, si des formules exactes n'existent pas ou si leur élaboration mathématique est trop difficile. Nous allons examiner dans cette section cinq procédures et expliquer leur emploi en termes de moyennes d'échantillon et de coefficients de régression. Dans tous les cas on ne fera pas intervenir le correcteur de population finie, puisque l'on retient par hypothèse un rapport d'échantillonnage peu élevé.

Si l'on pose par hypothèse un échantillon aléatoire simple, l'estimateur

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

peut servir à estimer la moyenne d'une variable X, où n est la taille de l'échantillon, x_j, $j = 1 \ldots n$ sont les observations. Un estimateur correspondant de la variance est l'expression bien connue

$$v(\bar{x}) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

Cette quantité a des propriétés bien connues et sans surprises. Il s'agit d'un estimateur sans biais de la variance vraie de l'échantillon σ^2 qui tend asymptotiquement vers une distribution khi-carré à mesure que n augmente.

Malheureusement, cette expression n'est valide que pour un échantillon aléatoire simple. Dans la réalité, cependant, de nombreux échantillons sont stratifiés. La stratification rend invalides les deux expressions ci-dessus. On peut résoudre cette difficulté en employant des poids. Les poids sont des nombres associés à chaque observation et qui rendent compte de son influence sur l'ensemble

which reflect its influence on the total population. A larger weight indicates greater influence. In each case, the weight given to an observation is directly proportional to the population of the stratum and inversely proportional to the number of observations in the stratum.

The weights, here denoted by w_j, once calculated can be used in the estimation procedures. The estimator for the mean from a stratified simple random sample becomes

$$\bar{x}_w = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$$

An estimator of the variance of x becomes

$$v(x) = \frac{\sum_{j=1}^{n} w_j (x_j - \bar{x}_w)^2}{\sum_{j=1}^{n} w_j}$$

and the variance of a sample mean becomes

$$v(\bar{x}_w) = \frac{(\sum_{j=1}^{n} w_j^2)(\sum_{j=1}^{n} w_j (x_j - \bar{x}_w)^2)}{(\sum_{j=1}^{n} w_j)^3}.$$

The Traditional Approach

The formulae for weighted variance estimation given above, while taking into account stratification, ignore the clustering used in many sample designs. Clustering destroys the assumption of independence of observations that both the simple random sample and stratified random sample schemes described above require. When clustered samples have been drawn, the variance estimators must become more sophisticated. A slight digression into survey design is needed before proceeding.

The 1974 Survey of Family Expenditures was essentially a stratified two-stage design. The first-stage units were segments (or clusters) chosen inside each of the cities of the survey and at the second stage households

de la population. Un poids plus élevé signifie une influence plus grande. Dans chaque cas, le poids attribué à chaque observation est directement proportionnel à la population de la strate et inversement proportionnel au nombre d'observations qu'elle contient.

Les poids, dénotés ici par w_j, une fois calculés peuvent servir aux procédures d'estimation. L'estimateur de la moyenne d'un échantillon aléatoire simple stratifié devient

$$\bar{x}_w = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$$

Un estimateur de la variance de x devient:

$$v(x) = \frac{\sum_{j=1}^{n} w_j (x_j - \bar{x}_w)^2}{\sum_{j=1}^{n} w_j}$$

et la variance d'une moyenne d'échantillon devient

$$v(\bar{x}_w) = \frac{(\sum_{j=1}^{n} w_j^2)(\sum_{j=1}^{n} w_j (x_j - \bar{x}_w)^2)}{(\sum_{j=1}^{n} w_j)^3}.$$

L'approche traditionnelle

Même si elles prennent en compte la stratification, les formules de l'estimation de la variance pondérée données ci-dessus ne tiennent pas compte du groupage en grappes de nombreux plans de sondage. Le groupage en grappes élimine l'hypothèse de l'indépendance des observations que les plans de sondage aléatoire simple et de sondage aléatoire stratifié décrits plus haut nécessitent. Une fois les échantillons en grappes tirés, les estimateurs de la variance doivent être plus élaborés. Avant de poursuivre, toutefois, une petite digression dans le domaine de l'élaboration des enquêtes s'impose.

L'enquête sur les dépenses des familles de 1974 utilisait essentiellement un sondage stratifié à deux degrés. Les unités du premier degré étaient des segments (ou grappes) choisis à l'intérieur de chacune des villes de l'enquête, et au deuxième

were chosen. For variance estimation purposes, the design strata were collapsed and denoted as paired areas. Each paired area was split into two components by alternately assigning segments of a paired area to one component or its complement.

The estimation of sample means is the same as it was in the case of a stratified sample, but the variance estimation is quite different. The formula used is that of the variance of a ratio estimator given, for example, in Kish(4). The sample mean can be expressed in other terms.

degré, les unités choisies étaient les ménages. Aux fins d'estimation de la variance, les strates du schéma ont été fondues et notées comme des régions couplées. Chaque région couplée a été partagée en deux composantes en affectant par alternance des segments d'une région couplée à une composante ou à son complément.

L'estimation des moyennes d'échantillon est la même que dans le cas d'un échantillon stratifié, mais l'estimation de la variance est assez différente. La formule utilisée est celle de la variance d'un estimateur de ratio donné, par exemple, dans Kish(4). La moyenne de l'échantillon peut être exprimée autrement.

$$\bar{x}_w = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j} = \frac{\sum_{j=1}^{L} \left(\sum_{j \in i_1} w_j x_j + \sum_{j \in i_2} w_j x_j \right)}{\sum_{j=1}^{L} \left(\sum_{j \in i_1} w_j + \sum_{j \in i_2} w_j \right)} = \frac{\dot{x}}{\dot{w}}$$

Now

Or

$$v(\bar{x}_w) = v\left(\frac{\dot{x}}{\dot{w}}\right) = \bar{x}_w^2 \left(\frac{v(\dot{x})}{\dot{x}^2} + \frac{v(\dot{w})}{\dot{w}^2} - \frac{2\, cov(\dot{x}, \dot{w})}{\dot{x}\ \dot{w}} \right)$$

where

où

$$v(\dot{x}) = \sum_{i=1}^{L} \left(\sum_{j \in i_1} w_j x_j - \sum_{j \in i_2} w_j x_j \right)^2 = \sum_{i=1}^{L} (Dx_i)^2$$

$$v(\dot{w}) = \sum_{i=1}^{L} \left(\sum_{j \in i_1} w_j - \sum_{j \in i_2} w_j \right)^2 = \sum_{i=1}^{L} (Dw_i)^2$$

$$cov(\dot{x}, \dot{w}) = \sum_{i=1}^{L} (Dx_i)(Dw_i)$$

and where L refers to the number of paired areas. The i's denote the paired areas and i_1 refers to component one and i_2 to component two. The estimate of standard error is obtained by taking the square root of the variance estimator. The variance of $\dot{x}$ and $\dot{w}$ are produced by essentially Keyfitz type estimators. The assumption of independence of weighting and survey results is dropped. A reader is referred to Keyfitz(3) for a more complete description.

et où L se rapporte au nombre de régions couplées. Les i dénotent les régions couplées, et i_1 se rapporte à la composante un, et i_2 à la composante deux. On obtient l'estimation de l'erreur-type en prenant la racine carrée de l'estimateur de la variance. On obtient la variance de $\dot{x}$ et de $\dot{w}$ essentiellement grâce à des estimateurs du type Keyfitz. On abandonne l'hypothèse de l'indépendance des poids et des résultats de l'enquête. Pour une description plus complète, le lecteur est prié de se reporter à Keyfitz(3).

Balanced Repeated Replications

The Keyfitz type estimator, although it considers the sampling design correctly, is restricted in its application. For determining standard errors of regression coefficients one scheme would be to determine separate regression estimates on each component of the paired areas and apply an estimator for standard errors similar to the above to the coefficient estimates. This is not appropriate with family expenditure data because the small number of observations in each paired area component precludes the estimation of the required coefficients. The necessity of producing satisfactory standard errors requires other techniques. The balanced repeated replication, and jackknife repeated replication techniques are two of these.

The balanced repeated replication procedure is discussed in Frankel(2), pages 34 to 39 and is outlined here. The technique is one of pseudo replication which is particularly useful when clustering destroys the independence of observations. Instead of replicating by taking a large number of surveys to determine variance estimates, a single survey is split into two half samples and the estimates from each are compared. Using the paired areas as defined in the previous section, half sample representation is considered by aggregating one component of each paired area. The complement half sample is obtained by considering the remaining components. These half samples have all the inherent properties of the full sample and can be used to produce any estimate that the whole sample will produce, but will do so only with less precision.

Replication is carried out by changing the composition of the half samples. In every case, the component of a given paired area and its complement will be in the opposite half samples. The number of replicates possible is determined by the number of paired areas.

At this point a more general piece of notation is introduced. Allow G(x) to generally denote the results of an estimation procedure carried out over the sample observations x. The G(x) can be weighted means, regression coefficients or any other statistic of interest.

If x_{hp} denotes the observations from half sample p, and x_{cp} its complement, the following four estimators of variance can be used.

Subdivision répétée équilibrée

Bien qu'il saisisse correctement, le plan de sondage, l'estimateur de type Keyfitz a des applications limitées. Pour la détermination des erreurs-types des coefficients de régression, on pourrait déterminer des estimations de régression séparées pour chaque composante des régions couplées et appliquer un estimateur des erreurs-types semblables à ceux ci-dessus aux estimations des coefficients. Cela n'est pas possible avec les données sur les dépenses des familles en raison du petit nombre d'observations dans chaque composante des régions couplées, qui empêche l'estimation des coefficients cherchés. Pour obtenir des erreurs-types satisfaisantes, il faut agir autrement. La technique de la subdivision répétée équilibrée et celle du jackknife sont deux solutions possibles.

La méthode de la subdivision répétée équilibrée est examinée par Frankel(2), pages 34 à 39, et nous la décrirons brièvement ici. Elle consiste en une pseudo-subdivision, qui est particulièrement utile lorsque la mise en grappes élimine l'indépendance des observations. Au lieu de subdiviser en prenant un nombre important d'enquêtes afin de déterminer les estimations de la variance, on partage une enquête unique en deux demi-échantillons et on compare les estimations de chaque demi-échantillon ainsi obtenu. En prenant les régions couplées définies dans la section précédente, on passe à la représentation par demi-échantillon en agrégeant une composante de chaque région couplée. On obtient le demi-échantillon complémentaire en prenant les composantes restantes. Ces demi-échantillons ont toutes les propriétés inhérentes à l'échantillon complet, et on peut s'en servir pour obtenir toute estimation que l'échantillon complet va produire, mais au détriment de la précision.

La subdivision se fait en changeant la composition des demi-échantillons. Dans chaque cas la composante d'une région couplée donnée et son complément seront dans le demi-échantillon opposé. Le nombre de subdivisions possibles est déterminé par le nombre de régions couplées.

Introduisons maintenant un élément de notation plus général. Soit G(x) la notation générale des résultats d'une procédure d'estimation faite sur les observations de l'échantillon. Les G(x) peuvent être des moyennes pondérées, des coefficients de régression ou toute autre statistique intéressante.

Si x_{hp} dénote les observations du demi-échantillon p, et si x_{cp} est son complément, on peut se servir des quatre estimations de la variance suivante:

$$v_1(G(x)) = \frac{1}{K} \sum_{p=1}^{K} (G(x_{hp}) - G(x))^2 \qquad (1)$$

$$v_2(G(x)) = \frac{1}{K} \sum_{p=1}^{K} (G(x_{cp}) - G(x))^2 \qquad (2)$$

$$v_3(G(x)) = \frac{1}{2} (v_1(G(x)) + v_2(G(x))) \qquad (3)$$

$$v(G(x)) = \frac{1}{4K} \sum_{p=1}^{K} (G(x_{hp}) - G(x_{cp}))^2 \qquad (4)$$

where K is the number of replicates used, and x denotes the observations from the full sample. The corresponding square roots yield the estimates of the standard error.

Two fairly obvious considerations are apparent with this sort of analysis. The first is that there are numerous ways of choosing the half sample to satisfy the above criterion of always splitting the two parts of a paired area. This is illustrated by observing that with L paired areas there are 2^{L-1} possible pairs of half samples. As the 1974 Family Expenditure Survey used 81 paired areas, there are 2^{80} different pairs of half samples. This is in the order of 10^{24}. The second consideration is that the estimate of the variances produced will depend, to some extent, on the particular half samples used for the estimate.

It is to these two problems that the balanced repeated replication method addresses itself. In order to reduce the effect of the choice of half paired areas in each half sample, replication may be performed by changing each time the composition of the half samples. Clearly, in our case, it is not feasible to produce every possible half sample. By repeating the estimation procedure for the half sample and its complement K times where K is the smallest integer divisible by 4 which is greater than the number of paired areas, a balanced procedure can be achieved. It has been shown that it is possible to put each paired area component in the same half sample as each component of every other paired area for exactly half the replications.

In our case, 84 pairs of half samples were determined since we had 81 paired areas in our sample.

$$v_1(G(x)) = \frac{1}{K} \sum_{p=1}^{K} (G(x_{hp}) - G(x))^2 \qquad (1)$$

$$v_2(G(x)) = \frac{1}{K} \sum_{p=1}^{K} (G(x_{cp}) - G(x))^2 \qquad (2)$$

$$v_3(G(x)) = \frac{1}{2} (v_1(G(x)) + v_2(G(x))) \qquad (3)$$

$$v(G(x)) = \frac{1}{4K} \sum_{p=1}^{K} (G(x_{hp}) - G(x_{cp}))^2 \qquad (4)$$

où K est le nombre de subdivisions utilisées et x dénote les observations de l'échantillon complet. Les racines carrées correspondantes donnent les estimations de l'erreur-type.

Ce type d'analyse fait ressortir deux considérations assez évidentes. La première est qu'il y a de nombreuses façons de choisir le demi-échantillon afin de répondre au critère ci-dessus du partage obligatoire des deux parties d'une région couplée. On peut s'en rendre compte si l'on remarque qu'avec L régions couplées, il y a 2^{L-1} couples possibles de demi-échantillons. Comme l'enquête sur les dépenses des familles de 1974 utilise 81 régions couplées, il y a 2^{80} différents couples de demi-échantillons. Un tel chiffre est de l'ordre de 10^{24}. La deuxième considération est que l'estimation de la variance obtenue va dépendre dans une certaine mesure des demi-échantillons particuliers utilisés pour l'estimation.

Ce sont ces deux problèmes que la méthode de la subdivision répétée équilibrée va traiter. Afin de réduire l'effet de la sélection de demi-régions couplées dans chaque moitié, il est possible d'effectuer une subdivision de l'échantillon en changeant chaque fois la composition des demi-échantillons. Il est clair que dans notre cas il n'est pas possible de produire tous les demi-échantillons possibles. En répétant la procédure d'estimation pour le demi-échantillon et son complément K fois, où K est le plus petit nombre entier divisible par 4 qui soit supérieur au nombre de régions couplées, on peut obtenir une procédure équilibrée. On a démontré qu'il est possible de placer chaque composante de région couplée dans le même demi-échantillon que chaque composante de toute autre région couplée pour exactement la moitié des subdivisions.

Dans notre cas, on a déterminé 84 paires de demi-échantillons, puisque nous avions 81 régions couplées dans notre échantillon.

To determine into which half sample a paired area component falls, an 84 by 81 matrix was constructed with entries A_{pi} equal to +1 or -1 such that

$$\sum_{p=1}^{84} A_{pi} A_{pi'} = 0, \text{ if } i \neq i'$$

In the p'th replicate if $A_{pi} = 1$ the first component of paired area i is included in half sample 1 and the complement component is included in half sample 2; if $A_{pi} = -1$ the reverse is true.

An explanation of the method of constructing these matrices is found in Plackett and Burman(7).

Once the sample has been delineated in this fashion, the statistics 1 to 4 can be used as estimators of the variance and their square roots for the standard error.

A more complete discussion is given in McCarthy(6).

Jackknife Techniques

Balanced repeated replication techniques provide one method of producing estimates of variances. The jackknife repeated replication procedure provides another. This method is also outlined in Frankel(2) pages 40 and 41 and is described here.

The jackknife procedure is also based on pseudo replications. In this technique, however, a replicate is formed by removing from the full sample one component of a paired area and including twice its complement. Similarly, the complement replicate is formed by removing the complement component and including twice the original component. The replicate formed by sample observations with a given paired area component replaced by its complement replicate is denoted by x_{ri} and the complement x_{ti}. Replication is performed by changing the paired area which is treated as above until all of the paired areas have been considered.

Four alternative estimates may be calculated:

$$v(G(x)) = \sum_{i=1}^{L} (G(x_{ri}) - G(x))^2 \quad (5)$$

Afin de déterminer dans quel demi-échantillon une composante de région couplée va se trouver, on a construit une matrice 84 x 81 dont les éléments A_{pi} sont égaux à +1 ou -1 de sorte que

$$\sum_{p=1}^{84} A_{pi} A_{pi'} = 0, \text{ if } i \neq i'$$

Dans la subdivision p'th, si $A_{pi} = 1$, la première composante de la région couplée i est comprise dans le demi-échantillon 1, et la composante complémentaire est comprise dans le demi-échantillon 2; si $A_{pi} = -1$, l'inverse est vrai.

On peut trouver dans Plackett et Burman(7) une explication de la façon de construire ces matrices.

Une fois l'échantillon défini de cette façon, les statistiques 1 à 4 peuvent servir d'estimateurs de la variance, et leurs racines carrées comme erreur-type.

La question est traitée de façon plus complète dans McCarthy(6).

Les techniques du jackknife

Les techniques de la subdivision répétée équilibrée sont une façon de produire des estimations de la variance. La technique du jackknife en est une autre. Frankel(2) donne aux pages 40 et 41 une brève description de cette méthode; nous allons en donner une aussi ci-dessous.

La technique du jackknife est également basée sur les pseudo-subdivisions. Toutefois, on forme alors une subdivision en enlevant de l'échantillon complet une composante d'une région couplée et en ajoutant deux fois son complément. De même, on forme la subdivision complémentaire en enlevant la composante complémentaire et en incluant deux fois la composante originale. La subdivision formée par les observations de l'échantillon avec une composante d'une région couplée donnée remplacée par sa subdivision complémentaire est dénotée x_{ri}, et le complément, x_{ti}. On effectue la subdivision en changeant la région couplée, qui est traitée comme on vient de l'exposer, jusqu'à ce que toutes les régions couplées aient été prises en compte.

On peut calculer quatre autres estimations:

$$v(G(x)) = \sum_{i=1}^{L} (G(x_{ri}) - G(x))^2 \quad (5)$$

$$v_2\,(G(x)) = \sum_{i=1}^{L} (G(x_{ti}) - G(x))^2 \qquad (6)$$

$$v_3\,(G(x)) = \frac{1}{2}\,(v_1(G(x)) + v_2(G(x))) \qquad (7)$$

$$v_4\,(G(x)) = \frac{1}{4} \sum_{i=1}^{L} (G\,(x_{ri}) - G(x_{ti}))^2 \qquad (8)$$

where G is a function as above. The standard error estimates required are the square roots of the respective variances.

The jackknife method has practical advantages over the balanced repeated replication scheme. In particular, there is no need to produce a large matrix $\underline{A}$ as in the case of the balanced repeated replication. It is therefore easier to write computer algorithms to produce the needed statistics.

Regression Coefficients

The standard errors of linear regression coefficients have traditionally been calculated using either standard non-weighted or weighted least squares procedures. The theory of regression is well known. Regression coefficients are calculated from the familiar formula

$$\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}$$

where $\hat{\underline{\beta}}$ is an fxl estimated coefficient vector, $\underline{X}$ is the nxf matrix of independent variables and $\underline{Y}$ the nxl dependent variable vector. If non-weighted least squares is used (that is, assuming a simple random sample), $\hat{\underline{\beta}}$ is chosen to minimize the expression

$$\sum_{i=1}^{n} \left(y_i - \sum_{m=1}^{f} \hat{\beta}_m\, x_{im}\right)^2$$

The variance estimates of the coefficients are given by

$$v(\hat{\beta}) = (\underline{X}'\underline{X})^{-1}\hat{\sigma}^{2}\ \star$$

$\star$ This expression in fact also provides estimates of the covariance between the parameter estimates in addition to the estimates of the respective variances.

The on-diagonal terms of the matrix $(\underline{X}'\underline{X})\hat{\sigma}^2$ provide the variances which are required. A further explanation is provided in most regression texts, for example, Draper and Smith(1).

$$v_2\,(G(x)) = \sum_{i=1}^{L} (G(x_{ti}) - G(x))^2 \qquad (6)$$

$$v_3\,(G(x)) = \frac{1}{2}\,(v_1(G(x)) + v_2(G(x))) \qquad (7)$$

$$v_4\,(G(x)) = \frac{1}{4} \sum_{i=1}^{L} (G\,(x_{ri}) - G(x_{ti}))^2 \qquad (8)$$

où G est une fonction comme ci-dessus. Les estimations de l'erreur-type cherchées sont les racines carrées des variances respectives.

La technique du jackknife a des avantages pratiques sur celle de la subdivision répétée équilibrée. En particulier, il n'est pas nécessaire de construire une grande matrice $\underline{A}$. Il est par conséquent plus facile d'écrire des algorithmes informatiques pour obtenir les statistiques désirées.

Coefficients de régression

Les erreurs-types des coefficients de régression linéaires sont habituellement calculées en utilisant la méthode des moindres carrés ordinaires, pondérés ou non. La théorie de la régression est bien connue. Les coefficients de régression sont calculés d'après la formule familière

$$\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}$$

où $\hat{\underline{\beta}}$ est un vecteur fxl des coefficients estimés, $\underline{X}$ est la matrice nxf des variables indépendantes et $\underline{Y}$ est le vecteur nxl des variables dépendantes. Si l'on se sert des moindres carrés non pondérés, c'est-à-dire si l'on suppose un échantillon aléatoire simple, on choisit $\hat{\underline{\beta}}$ de façon à minimiser l'expression

$$\sum_{i=1}^{n} \left(y_i - \sum_{m=1}^{f} \hat{\beta}_m\, x_{im}\right)^2$$

Les estimations de la variance des coefficients sont données par

$$v(\hat{\beta}) = (\underline{X}'\underline{X})^{-1}\hat{\sigma}^{2}\ \star$$

$\star$ Cette expression donne également en fait les estimations de la covariance entre les estimations des paramètres, en plus des estimations des variances respectives. Les termes diagonaux de la matrice $(\underline{X}'\underline{X})\,\hat{\sigma}^2$ donnent les variances cherchées. On peut trouver d'autres explications dans la plupart des textes traitant de la régression, tels Draper et Smith(1).

where

$$\hat{\sigma}^2 = \frac{1}{n-f} \sum_{i=1}^{n} (y_i - \sum_{m=1}^{f} \hat{\beta}_m x_{im})^2$$

Unweighted least square makes no allowance for stratification in the sampling procedure. For this purpose, weighted least squares must be used. The vector $\hat{\underline{\beta}}$ is then chosen to minimize

$$\sum_{i=1}^{n} w_i (y_i - \sum_{m=1}^{f} \hat{\beta}_m x_{im})^2$$

Similar formulae to the above which provide estimates of the coefficients and their standard errors can be applied. The vector $\underline{Y}$ and the matrix $\underline{X}$ are replaced by the corresponding $\underline{W}\,\underline{Y}$ and $\underline{W}\,\underline{X}$ respectively where $\underline{W}$ is the nxn matrix with the ith diagonal entry equal to $\sqrt{w_i}$, the square root of the weight associated with observation i, and zeros elsewhere, $\hat{\sigma}^2$ becomes in this instance

$$\frac{1}{n-f} \sum_{i=1}^{n} w_i (y_i - \sum_{m=1}^{f} \hat{\beta}_m x_{im})^2$$

For a more complete description of regression analysis, the reader is referred to Draper and Smith(1).

Although weighted least squares techniques produce unbiased regression coefficients, the clustering in many sample designs still gives rise to problems for variance estimation. Keyfitz estimators are not normally used for this because the production of separate estimates for each component of a paired area is not practical given the small sample sizes within the paired areas.

Balanced repeated replication does not suffer from this sort of restriction, however. In a half sample there are enough observations to produce stable coefficient estimates so this procedure can produce needed standard errors. Separate $\underline{X}'\underline{X}$ and $\underline{X}'\underline{Y}$ matrices are computed for each of the half samples, determined using a scheme similar to that described previously. Regression coefficients are calculated from these and the estimators for the variance of the coefficients are given by formulae similar to (1) to (4). In this report the only estimates tabulated are those for (4):

où

$$\hat{\sigma}^2 = \frac{1}{n-f} \sum_{i=1}^{n} (y_i - \sum_{m=1}^{f} \hat{\beta}_m x_{im})^2$$

Les moindres carrés non pondérés ne prennent pas en compte la stratification lors de l'échantillonnage. Il faut donc utiliser les moindres carrés pondérés. On choisit le vecteur $\hat{\underline{\beta}}$ de façon à minimiser

$$\sum_{i=1}^{n} w_i (y_i - \sum_{m=1}^{f} \hat{\beta}_m x_{im})^2$$

Des formules semblables à celle-ci et donnant des estimations des coefficients et de leurs erreurs-types peuvent être utilisées. Le vecteur $\underline{Y}$ et la matrice $\underline{X}$ sont remplacés par les $\underline{W}\,\underline{Y}$ et $\underline{W}\,\underline{X}$ respectivement correspondants, où $\underline{W}$ est la matrice nxn dont le i-ième élément de la diagonale est égal à $\sqrt{w_i}$, la racine carrée du poids associé à l'observation i, et des zéros ailleurs; $\hat{\sigma}^2$ devient dans ce cas

$$\frac{1}{n-f} \sum_{i=1}^{n} w_i (y_i - \sum_{m=1}^{f} \hat{\beta}_m x_{im})^2$$

Pour une analyse plus complète de la régression, le lecteur est prié de se reporter à Draper et Smith(1).

Bien que les techniques des moindres carrés pondérés donnent des coefficients de régression sans biais, le groupage en grappes dans de nombreux plans de sondage va quand même poser des problèmes pour l'estimation de la variance. On ne se sert pas habituellement des estimateurs de Keyfitz, car l'obtention d'estimations distinctes pour chaque composante d'une région couplée n'est pas pratique en raison de la petite taille des échantillons au sein des régions couplées.

La subdivision répétée équilibrée ne souffre pas de cet inconvénient. Il y a assez d'observations dans un demi-échantillon pour obtenir des estimations stables des coefficients; autrement dit, cette méthode permet d'obtenir les erreurs-types voulues. On calcule des matrices $\underline{X}'\underline{X}$ et $\underline{X}'\underline{Y}$ distinctes pour chacun des demi-échantillons qui ont été déterminés selon une méthode semblable à celle décrite précédemment. On calcule à partir de ces matrices les coefficients de régression, et l'estimateur de la variance des coefficients est donné par des formules semblables aux formules (1) à (4). Ici, les seules estimations totalisées sont celles pour (4):

$$v(\hat{\underline{\beta}}) = \frac{1}{4K} \sum_{p=1}^{K} (G(x_{hp}) - G(x_{cp}))^2$$

The variable K is as previously defined but in this instance $G(x_{hp})$ and $G(x_{cp})$ are the values of the coefficients calculated from the half sample and its complement respectively.

The main potential problems with this sort of analysis would occur if the half samples did not have enough records to produce estimates with some form of statistical stability or if the $\underline{X}'\underline{X}$ matrices were singular. Due to the technique of assigning records to paired areas and of subsequently constructing the half samples, the former problem is not considered to be troublesome. An investigation of the results in this study revealed no evidence of the latter.

The appeal of using balanced repeated replication is that one only has to apply the regression procedure to each half sample followed by a straightforward variance computation. The resulting scheme is only an approximation of a valid procedure. Moreover, the exact characteristics of the variance estimates produced are not known. Again, the reader is referred to McCarthy. It has been, nevertheless, suggested by Kish and Frankel(5) that the estimates of regression coefficients based on a clustered sample do, in fact, tend towards normality about the appropriate parameters, but that the approach to normality is retarded by the correlations between the clusters.

The jackknife procedure is also applicable to regression analysis. In this case, $\underline{X}'\underline{X}$ and $\underline{X}'\underline{Y}$ matrices are produced for each replicate and their complements. The replicates are chosen as previously, by the inclusion of twice the complement component of a paired area and the deletion of the first paired area component. The complement replication is formed by producing $\underline{X}'\underline{X}$ and $\underline{X}'\underline{Y}$ matrices for the whole sample with the first component of a paired area included twice and the complement component dropped. After the replication matrices have been produced the regression estimates can be calculated and statistics (5) to (8) determined. We have used only the estimates which correspond to (8).

$$v(\hat{\underline{\beta}}) = \frac{1}{4} \sum_{i=1}^{L} (G(x_{ri}) - G(x_{ti}))^2$$

$$v(\hat{\underline{\beta}}) = \frac{1}{4K} \sum_{p=1}^{K} (G(x_{hp}) - G(x_{cp}))^2$$

La variable K est définie comme précédemment, sauf que ici $G(x_{hp})$ et $G(x_{cp})$ sont les valeurs des coefficients calculés à partir du demi-échantillon et de son complément respectivement.

Les principaux problèmes potentiels reliés à ce type d'analyse se présenteraient si les demi-échantillons n'avaient pas assez d'éléments pour produire des estimations ayant une certaine forme de stabilité statistique, ou si les matrices $\underline{X}'\underline{X}$ étaient singulières. En raison de la technique qui consiste à affecter à chaque région couplée les dossiers et ensuite à construire les demi-échantillons, le problème précédent n'est pas considéré comme préoccupant. Une étude des résultats, dans le cas de la présente étude, n'a rien révélé.

L'intérêt de la technique de la subdivision répétée équilibrée est qu'il faut simplement appliquer la méthode de régression à chaque demi-échantillon, et calculer ensuite directement la variance. Le schéma qui en résulte n'est qu'une approximation d'une procédure valide. De plus, les caractéristiques exactes des estimations de la variance ne sont pas connues. Là encore, le lecteur est prié de se reporter à McCarthy. Kish et Frankel(5) ont néanmoins suggéré que les estimations des coefficients de régression établis sur la base d'un échantillon en grappes tendent en fait vers la normalité pour ce qui est des paramètres appropriés, mais que cette tendance à la normalité est freinée par les corrélations entre les grappes.

On peut aussi appliquer la technique du jackknife à l'analyse de régression. Les matrices $\underline{X}'\underline{X}$ et $\underline{X}'\underline{Y}$ sont alors produites pour chaque subdivision et leurs compléments. Les subdivisions sont choisies comme auparavant, par l'inclusion de deux fois la composante complémentaire d'une région couplée et par la suppression de la première composante de la région couplée. La subdivision complémentaire se fait en produisant les matrices $\underline{X}'\underline{X}$ et $\underline{X}'\underline{Y}$ pour la totalité de l'échantillon, la première composante d'une région couplée étant comprise deux fois et la composante complémentaire abandonnée. Une fois les matrices de subdivision produites, on peut calculer les estimations de régression et déterminer les statistiques (5) à (8). Seules les estimations correspondant à (8) sont utilisées.

$$v(\hat{\underline{\beta}}) = \frac{1}{4} \sum_{i=1}^{L} (G(x_{ri}) - G(x_{ti}))^2$$

As with balanced repeated replication, the properties of the estimates produced by the jackknife technique are not well known. The reader is referred to Frankel for a discussion of this problem.

Comme pour la méthode de la subdivision répétée équilibrée, les propriétés des estimations obtenues par la technique du jackknife ne sont pas bien connues. Pour un examen de ce problème, le lecteur voudra bien se reporter à Frankel.

SECTION III

THE RESULTS

Sample Means

In Section II of this report, various estimators of the standard errors of sample means and of regression coefficients were described. In this section, the results of these estimators will be presented using data from the 1974 Survey of Family Expenditures. In all cases, the finite population correction has not been used. This was done since the proportion of the population sampled was quite small.

Given the delineation of paired areas, the variance estimators of linear statistics produced by the balanced repeated replication, the jackknife and traditional Keyfitz methods give identical results. For the sample means, which are ratio estimates in which the zero correlation between weights and results assumption is dropped, the results are almost identical, and consequently, only the traditional results are shown in the tables.

The results obtained from the stratified estimator are, in general, considerably smaller than the traditional estimates. The ratio of traditional to stratified estimator is commonly of the order of 1.0 to 1.3 but is considerably higher in some cases. This difference is most probably due to the deviation of the sample design from stratified random to two-stage clustered. The clustering results in destruction of the independence of observations assumption which effectively reduces the sample size as far as variance estimation is concerned. Evidence of this can be obtained from an examination of Text Table I. These results justify the concern over the correct procedure and invalidate the use of the stratified standard error estimates on clustered samples for the production of statistics such as coefficients of variation and for statistical tests of significance.

SECTION III

LES RÉSULTATS

Moyennes de l'échantillon

Nous avons décrit dans la Section II divers estimateurs des erreurs-types des moyennes des échantillons et des coefficients de régression. Nous allons présenter ici les résultats de ces estimateurs en nous servant des données provenant de l'enquête sur les dépenses des familles de 1974. Dans tous les cas, on n'a pas utilisé la correction pour la population finie, puisque la proportion de la population échantillonnée était assez petite.

Compte tenu de la délimitation des régions couplées, les estimateurs de la variance des statistiques linéaires obtenues par la méthode de la subdivision répétée équilibrée, celle du jackknife et la méthode traditionnelle de Keyfitz ont donné des résultats identiques. Dans le cas des moyennes de l'échantillon, qui sont des estimations des ratios pour lesquelles on abandonne l'hypothèse de corrélation nulle entre les poids et les résultats, les résultats sont presque identiques et par conséquent seuls les résultats traditionnels sont présentés dans les tableaux.

Les résultats obtenus grâce aux estimateurs stratifiés sont en général considérablement inférieurs aux résultats traditionnels. Le ratio estimateur traditionnel/estimateur stratifié est souvent de l'ordre de 1.0 à 1.3, mais il est considérablement supérieur dans certains cas. Cette différence s'explique presque certainement par le passage d'un plan de sondage aléatoire stratifié à un plan à deux degrés en grappes. La mise en grappes se traduit par l'élimination de l'hypothèse de l'indépendance des observations, qui dans les faits réduit la taille de l'échantillon pour ce qui est de l'estimation de la variance. Un examen du tableau explicatif I le prouve. Ces résultats justifient les préoccupations reliées à l'utilisation de la bonne procédure et invalident l'emploi des estimations des erreurs-types stratifiées dans les échantillons en grappes pour l'obtention de statistiques telles que les coefficients de variation, et pour les tests d'hypothèses.

TEXT TABLE A. Estimates of Standard Error of Mean

TABLEAU EXPLICATIF 1. Estimations de l'erreur-type de la moyenne

	Standard Error / Erreur-type			Ratio(1) Keyfitz to Stratified / Ratio(1) Keyfitz/ Stratif.
	Mean / Moyenne	Stratified / Stratification	Keyfitz type / Keyfitz	
Average - Moyenne				
Family size - Taille de la famille	2.98	0.03	0.03	
Number of children under 5 years - Nombre d'enfants de moins de 5 ans	0.23	0.01	0.01	
Number of children 5-15 years - Nombre d'enfants 5-15 ans	0.62	0.02	0.02	
Number of adults 16-17 years - Nombre d'adultes 16-17 ans	0.12	0.01	0.01	
Number of adults 18-64 years - Nombre d'adultes 18-64 ans	1.85	0.02	0.02	
Number of adults 65+ years - Nombre d'adultes 65 ans et plus	0.20	0.01	0.01	
Number of full-time earners - Nombre de gagneurs à plein temps	0.92	0.01	0.01	
Age of head - Âge du chef	44.75	0.3	0.3	
Net income before taxes - Revenu net avant impôt	14,510.4	153.3	177.0	1.2
Other money receipts - Autres rentrées monétaires	358.6	26.7	39.2	1.1
Net change in assets, liabilities - Variation nette des éléments d'actif et de passif	752.4	56.7	51.3	1.0
Percentage - Pourcentage				
Home-owners - Propriétaires de maison	50.3	0.79	1.37	1.7
Car and truck owners - Propriétaires de voitures et camions	71.4	0.71	1.06	1.5
With head Canadian born - Chef né Canadien	68.7	0.73	1.21	1.7
With wife employed full-time - Épouse employée à plein temps	16.3	0.58	0.75	1.3
Average dollar expenditure per family - Dépenses moyennes en dollars par famille				
Food - Alimentation	2,412.3	19.4	17.1	0.9
Shelter - Logement	2,102.2	19.6	31.8	1.6
Rented living quarters - Logement loué	769.0	14.3	27.5	1.9
Owned living quarters - Logement appartenant à l'occupant	904.2	20.3	40.9	2.0
Other housing - Autre logement	98.4	4.5	5.5	1.2
Water and fuel - Eau et combustible	330.5	4.2	6.2	1.5
Household operation - Entretien du ménage	526.6	7.2	8.3	1.1
Furnishing and equipment - Articles d'ameublement et accessoires	683.3	14.8	14.7	1.0
Household appliances - Appareils ménagers	148.7	4.4	5.4	1.2
Other - Autres	534.6	12.2	10.8	0.9
Clothing - Habillement	993.4	13.4	13.8	1.0
Personal care - Soins personnels	239.7	2.9	3.1	1.1
Medical and health care - Soins médicaux et de santé	297.6	4.5	5.7	1.3
Smoking and alcoholic beverages - Articles pour fumeurs et alcools	478.3	8.3	9.2	1.1
Travel and transportation - Déplacements et transport	1,712.0	31.7	39.9	1.3
Automobile (and truck) - Automobile (et camion)	1,337.4	28.1	33.9	1.2
Purchase - Achat	630.8	23.1	25.3	1.1
Operation - Utilisation	706.6	10.6	17.0	1.6
Other - Autres	374.6	12.8	14.4	1.1
Recreation - Loisirs	541.0	10.3	12.9	1.3
Reading - Lecture	81.4	1.5	2.0	1.4
Education - Éducation	107.0	4.5	5.5	1.2
Miscellaneous expenses - Dépenses diverses	292.3	6.9	6.6	1.0
Current consumption - Total - Dépenses de consommation courantes	10,467.0	87.4	104.9	1.2
Personal taxes - Impôts personnels	2,580.7	53.9	66.1	1.2
Security - Sécurité	706.6	14.5	18.0	1.2
Gifts and contributions - Dons et contributions	295.1	9.7	10.4	1.1
Expenditure - Total - Dépenses	14,049.4	137.6	170.8	1.2

(1) This ratio was not calculated for the first eight variables due to the lack of required significant figures. - Ce ratio n'a pas été calculé pour les huit premières variables en raison de l'absence des chiffres significatifs nécessaires.

Of further interest will be the distribution of the variance estimates. At this time, the form of the distribution is not known but could be determined by a further empirical study. The results could then be compared with both the distribution assuming simple random sampling and with results from other statistical surveys.

La distribution des estimations de la variance présentera encore plus d'intérêt. À l'heure actuelle, la forme de la distribution n'est pas connue, mais elle pourrait être déterminée par une étude empirique plus poussée. On pourrait alors comparer les résultats à la fois à ceux d'une distribution dans l'hypothèse d'un échantillonnage aléatoire simple et à ceux d'autres enquêtes statistiques.

Regression Coefficients

The results of using the balanced repeated replication and jackknife repeated replication methods to estimate standard errors of regression coefficients are found in Text Tables II, III and IV. The results are tabulated alongside those for weighted least squares. The dependent variables used were the three basic expenditures, food, shelter and clothing. The regression coefficients have been tabulated for completeness. The ratios of the standard errors of coefficients estimated by the jackknife method divided by those of weighted least squares have been tabulated along with the number of observations in the sample reporting an attribute, under the column heading sample size.

In the analysis of the 1969 Family Expenditure Survey, there was little evidence that the use of sample weights influenced the estimates of the regression coefficients. This probably does not hold true with respect to the variance estimates however. Traditionally, unweighted regressions were used as they facilitated comparisons of various partitions of the data base. In the case of 1974, however, it is expected that weighting will change the estimates of both the coefficients and the standard errors. This is caused by the low income part of the sample, which will give rise in the unweighted case to undue influence of the low income families. It is more appropriate, therefore, to use both weighted regressions and improved variance estimation techniques.

A linear model was used for the purpose of estimating expenditures. This model used total expenditure with dummy variables for family type, tenure type, age of head and city of residence as predictor variables. This linear model was chosen because previous work with family expenditure surveys from other years had shown that satisfactory performance could be expected. Dummy variables take a value of 0 or 1, a value 1 is taken if the family has the attribute in question, 0 is taken otherwise. The inversion of $\underline{X}'\underline{X}$ requires that there be no singularities in the data and, consequently, one variable from each

Coefficients de régression

Les résultats de l'utilisation de la subdivision répétée équilibrée et de la technique du jackknife pour estimer les erreurs-types des coefficients de régression se trouvent aux tableaux explicatifs II, III et IV. Ces tableaux contiennent également ceux obtenus par la méthode des moindres carrés pondérés. Les variables dépendantes ont été les trois dépenses fondamentales: aliments, logement et vêtements. On a totalisé les coefficients de régression pour des raisons d'exhaustivité. Les ratios des erreurs-types des coefficients estimés par la technique du jackknife divisés par ceux obtenus par la méthode des moindres carrés pondérés ont été totalisés à côté du nombre d'observations dans l'échantillon déclarant un caractère, en plus d'une colonne donnant la taille de l'échantillon.

Lors de l'analyse de l'enquête sur les dépenses des familles de 1969, il y avait peu de preuves que l'utilisation des poids des échantillons influençait les estimations des coefficients de régression. Ceci ne serait probablement pas vrai dans le cas de l'estimation de la variance, toutefois. Traditionnellement, on se sert de régressions non pondérées, car elles facilitent la comparaison de diverses partitions de la base de données. Dans le cas de l'enquête de 1974, cependant, on s'attend à ce que la pondération va changer les estimations des coefficients et des erreurs-types. La cause en est l'élément revenu inférieur de l'échantillon, qui, dans l'échantillon non pondéré, va donner une importance excessive aux familles à faibles revenus. Il est donc préférable d'utiliser à la fois des régressions pondérées et des techniques améliorées d'estimation de la variance.

On s'est servi d'un modèle linéaire pour estimer les dépenses. Ce modèle utilisait le total des dépenses avec des variables auxiliaires pour le type de familles, le mode d'occupation, l'âge du chef et la ville de résidence comme variables explicatives. On a retenu ce modèle parce que les travaux précédents portant sur les enquêtes sur les dépenses des familles des autres années ont montré qu'on pouvait s'attendre à des résultats satisfaisants. Les variables auxiliaires prennent des valeurs de 0 ou de 1; on retient la valeur 1 si la famille possède le caractère en question, 0 autrement. L'inversion de $\underline{X}'\underline{X}$ exige l'absence de singularité des données et il faut par conséquent

TEXT TABLE II. Regression Coefficients and Estimated Standard Errors Dependent Variable: Food Expenditure

TABLEAU EXPLICATIF II. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses en aliments

No.		Regression Coefficient / Coefficient de régression	Standard errors / Erreurs-types: Weighted least square / Moindres carrés pondérés	Standard errors / Erreurs-types: Balanced(1) repeated replication / Subdivision(1) répétée équilibrée
1	Total expenditure	0.06206	0.00128	0.00384
2	One adult, no children	- 464.35	29.29	39.09
3	Three adults, no children	393.28	40.45	53.73
4	Four adults, no children	942.39	51.90	90.18
5	One adult, one or more children	34.72	58.87	61.46
6	Two adults, one child	303.89	33.93	33.06
7	Two adults, two children	492.37	34.53	37.23
8	Two adults, three children	766.18	49.50	66.77
9	Two adults, four children	1,095.8	78.23	101.25
10	Two adults, five or more children	1,276.7	119.54	166.21
11	Three adults, one child	817.42	52.66	93.08
12	Three adults, two or more children	1,107.7	50.60	96.43
13	Other family types	1,496.6	42.91	71.07
14	Age of head less than 25 years	- 283.21	41.68	51.46
15	Age of head 25-34 years	- 247.10	30.52	48.91
16	Age of head 35-44 years	- 60.15	30.47	60.59
17	Age of head 55-64 years	- 131.82	32.67	75.5
18	Age of head 65 years or more	- 256.24	36.59	56.60
19	Home-owners with mortgage	- 118.26	29.32	42.54
20	Tenants: regular	- 57.87	28.86	46.05
21	Tenants: roomers	58.22	55.90	93.30
22	Tenants: rent free	- 19.91	91.17	103.82
23	Mixed tenure	- 249.65	53.03	65.84
24	St. John's 6(3)	33.12	94.83	39.69
25	Halifax 6	- 171.61	72.66	47.78
26	Saint John 3	- 27.90	94.12	44.91
27	Québec 7	237.11	55.12	47.41
28	Montréal 9	324.05	41.04	47.60
29	Ottawa 5	60.29	56.58	45.29
30	Toronto 9	180.72	40.45	61.90
31	Thunder Bay 3	51.19	92.51	51.78
32	Regina 3	- 198.32	80.61	92.53
33	Saskatoon 4	- 199.27	85.34	46.69
34	Calgary 6	- 45.88	55.94	41.86
35	Edmonton 6	- 71.60	53.63	36.04
36	Vancouver 8	135.84	44.62	59.97
37	Constant	1,307.2	53.07	86.71

(1) Estimators #4 for standard errors were used both for balanced repeated replication and jackknife repeated replication. See pages 15 and 16 of this report. The standard error ratio is given for the jackknife procedure estimates only.

(2) The sample sizes for the dummy variables give the number of observations which have the attribute:

Two adults, no children	1,690
Age of head 45-64	1,119
Home-owner, no mortgage	1,334
Winnipeg City	491

(3) These figures denote the number of paired areas for the cities. Winnipeg has six paired areas.

TEXT TABLE II. Regression Coefficients and Estimated Standard Errors Dependent Variable: Food Expenditure

TABLEAU EXPLICATIF II. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses en aliments

Standard errors Erreurs-types	Ratio jackknife to weighted	Sample(2) size		
Jackknife(1) repeated replicaton	Ratio jackknife/ M.C. pondérés	Taille de l'échan-tillon(2)		
Subdivision répétée jackknife(1)				N°
0.00372	2.9		Total, dépenses	1
38.68	1.3	1,408	Un adulte, sans enfants	2
52.95	1.3	396	Trois adultes, sans enfants	3
89.33	1.7	198	Quatre adultes, sans enfants	4
60.64	1.0	236	Un adulte, un enfant ou plus	5
32.86	1.0	683	Deux adultes, un enfant	6
37.01	1.1	732	Deux adultes, deux enfants	7
66.32	1.3	303	Deux adultes, trois enfants	8
99.45	1.3	98	Deux adultes, quatre enfants	9
160.07	1.3	48	Deux adultes, cinq enfants ou plus	10
91.91	1.7	205	Trois adultes, un enfant	11
95.71	1.9	257	Trois adultes, deux enfants ou plus	12
70.40	1.6	376	Autres types de famille	13
51.19	1.2	548	Âge du chef: moins de 25 ans	14
48.39	1.6	1,569	Âge du chef: 25-34 ans	15
60.11	2.0	1,234	Âge du chef: 35-44 ans	16
74.73	2.3	930	Âge du chef: 55-64 ans	17
56.25	1.3	1,230	Âge du chef: 65 ans ou plus	18
42.05	1.4	1,809	Propriétaires avec hypothèque	19
45.52	1.2	2,954	Locataires: ordinaires	20
89.56	1.6	188	Locataires: chambreurs	21
101.74	1.1	84	Locataires: sans loyer	22
64.75	1.2	261	Mode d'occupation mixte	23
39.39	0.4	355	St. John's 6(3)	24
47.50	0.7	371	Halifax 6	25
43.92	0.5	269	Saint John 3	26
46.87	0.9	469	Québec 7	27
47.11	1.1	1,000	Montréal 9	28
45.10	0.8	350	Ottawa 5	29
60.82	1.5	813	Toronto 9	30
51.02	0.6	315	Thunder Bay 3	31
91.91	1.1	341	Regina 3	32
46.35	0.5	353	Saskatoon 4	33
41.56	0.7	365	Calgary 6	34
35.65	0.7	518	Edmonton 6	35
59.42	1.3	620	Vancouver 8	36
85.31	1.6		Constante	37

(1) Les estimateurs #4 des erreurs-types ont été utilisés pour les deux types de subdivision. Voir les pages 15 et 16 du présent document. Le ratio de l'erreur-type n'est donné que pour les estimations obtenues par la méthode du jackknife seulement.

(2) Les tailles d'échantillon pour les variables auxiliaires donnent le nombre d'observations qui possèdent le caractère:

Deux adultes, sans enfants	1,690
Âge du chef: 45-64 ans	1,119
Propriétaires, pas d'hypothèque	1,334
Winnipeg	491

(3) Ces chiffres dénotent le nombre de régions couplées pour les villes. Winnipeg a six régions couplées.

TEXT TABLE III. Regression Coefficients and Estimated Standard Errors Dependent Variables: Shelter Expenditure

TABLEAU EXPLICATIF III. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses pour l'hébergement

No.		Regression Coefficient / Coefficient de régression	Standard errors / Erreurs-types	
			Weighted least squares / Moindres carrés pondérés	Balanced(1) repeated replication / Subdivision(1) répétée équilibrée
1	Total expenditure	0.07363	0.00154	0.00666
2	One adult, no children	42.159	35.24	44.99
3	Three adults, no children	- 106.94	48.67	61.88
4	Four adults, no children	- 329.20	62.43	97.43
5	One adult, one or more children	217.36	70.82	89.56
6	Two adults, one child	51.13	40.82	46.67
7	Two adults, two children	123.90	41.54	75.83
8	Two adults, three children	67.44	59.31	84.96
9	Two adults, four children	84.71	94.11	114.43
10	Two adults, five or more children	159.15	143.82	407.05
11	Three adults, one child	- 127.56	63.35	67.24
12	Three adults, two or more children	- 152.57	60.88	91.15
13	Other family types	- 72.02	51.63	100.98
14	Age of head less than 25 years	34.96	50.15	65.40
15	Age of head 25-34 years	179.36	36.72	60.25
16	Age of head 35-44 years	122.44	36.66	50.00
17	Age of head 55-64 years	84.48	39.31	53.76
18	Age of head 65 years or more	197.41	44.02	53.84
19	Home-owners with mortgage	896.50	35.27	69.72
20	Tenants: regular	343.81	34.72	58.13
21	Tenants: roomers	- 261.73	67.25	97.58
22	Tenants: rent free	-1,077.3	109.69	89.01
23	Mixed tenure	1,338.1	63.79	121.48
24	St. John's (3)	320.15	114.09	108.27
25	Halifax	316.34	87.42	83.14
26	Saint John	249.17	113.24	97.83
27	Québec	96.99	67.04	98.22
28	Montréal	180.31	49.38	89.84
29	Ottawa	392.54	68.07	99.84
30	Toronto	411.57	48.67	95.20
31	Thunder Bay	- 4.41	111.30	81.16
32	Regina	- 126.42	96.99	79.71
33	Saskatoon	- 69.35	102.67	87.58
34	Calgary	36.73	67.30	96.08
35	Edmonton	47.43	64.53	78.94
36	Vancouver	387.91	53.69	82.57
37	Constant	261.91	63.85	108.17

See footnotes to Text Table II.

TEXT TABLE III. Regression Coefficients and Estimated Standard Errors Dependent Variable: Shelter Expenditure

TABLEAU EXPLICATIF III. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses pour l'hébergement

Standard errors Erreurs-types	Ratio jackknife to weighted	Sample(2) size		
Jackknife(1) repeated replicaton	Ratio jackknife/ M.C. pondérés	Taille de l'échan-tillon(2)		
Subdivision répétée jackknife(1)				N°
0.00652	4.2		Total, dépenses	1
44.57	1.3	1,408	Un adulte, sans enfants	2
61.24	1.3	396	Trois adultes, sans enfants	3
96.62	1.5	198	Quatre adultes, sans enfants	4
88.32	1.2	236	Un adulte, un enfant ou plus	5
46.57	1.1	683	Deux adultes, un enfant	6
75.47	1.8	732	Deux adultes, deux enfants	7
83.98	1.4	303	Deux adultes, trois enfants	8
112.15	1.2	98	Deux adultes, quatre enfants	9
387.86	2.7	48	Deux adultes, cinq enfants ou plus	10
66.27	1.0	205	Trois adultes, un enfant	11
90.51	1.5	257	Trois adultes, deux enfants ou plus	12
99.92	1.9	376	Autres types de famille	13
64.87	1.3	548	Âge du chef: moins de 25 ans	14
59.96	1.6	1,569	Âge du chef: 25-34 ans	15
49.67	1.4	1,234	Âge du chef: 35-44 ans	16
53.42	1.4	930	Âge du chef: 55-64 ans	17
53.70	1.2	1,230	Âge du chef: 65 ans ou plus	18
69.09	2.0	1,809	Propriétaires avec hypothèque	19
57.81	1.7	2,954	Locataires: ordinaires	20
96.57	1.4	188	Locataires: chambreurs	21
87.02	0.8	84	Locataires: sans loyer	22
120.26	1.9	261	Mode d'occupation mixte	23
108.12	0.9	355	St. John's(3)	24
82.97	0.9	371	Halifax	25
97.78	0.9	269	Saint John	26
98.01	1.5	469	Québec	27
89.54	1.8	1,000	Montréal	28
99.00	1.4	350	Ottawa	29
94.58	1.9	813	Toronto	30
80.79	0.7	315	Thunder Bay	31
79.21	0.8	341	Regina	32
87.29	0.9	353	Saskatoon	33
95.69	1.4	365	Calgary	34
78.75	1.2	518	Edmonton	35
82.47	1.5	620	Vancouver	36
107.12	1.7		Constante	37

Voir les renvois du tableau explicatif II.

TEXT TABLE IV. Regression Coefficients and Estimated Standard Errors Dependent Variable: Clothing Expenditure
TABLEAU EXPLICATIF IV. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses en vêtements

No.		Regression Coefficient Coefficient de régression	Standard errors Erreurs-types	
			Weighted least squares Moindres carrés pondérés	Balanced(1) repeated replication Subdivision(1) répétée équilibrée
1	Total expenditure	0.05779	0.00105	0.00627
2	One adult, no children	- 81.12	23.96	28.21
3	Three adults, no children	162.67	33.10	50.57
4	Four adults, no children	468.95	42.46	69.93
5	One adult, one or more children	71.66	48.17	52.99
6	Two adults, one child	104.58	27.76	34.58
7	Two adults, two children	112.95	28.25	26.65
8	Two adults, three children	164.74	40.34	47.07
9	Two adults, four children	150.27	64.01	71.90
10	Two adults, five or more children	- 98.50	97.81	311.65
11	Three adults, one child	296.04	43.09	75.70
12	Three adults, two or more children	408.73	41.40	72.72
13	Other family types	522.50	35.11	78.13
14	Age of head less than 25 years	20.06	34.01	51.93
15	Age of head 25-34 years	- 26.18	24.97	36.53
16	Age of head 35-44 years	- 15.54	24.93	40.48
17	Age of head 55-64 years	- 106.92	26.73	42.22
18	Age of head 65 years or more	- 98.83	29.94	48.03
19	Home-owners with mortgage	- 36.14	23.99	46.54
20	Tenants: regular	37.23	23.61	37.10
21	Tenants: roomers	15.12	45.74	45.14
22	Tenants: rent free	123.97	74.60	74.66
23	Mixed tenure	- 207.10	43.39	59.51
24	St. John's (3)	- 52.15	77.66	34.51
25	Halifax	- 21.39	59.45	33.26
26	Saint John	- 33.19	77.01	39.63
27	Québec	99.21	45.59	47.66
28	Montréal	55.96	33.58	46.37
29	Ottawa	51.84	46.30	36.66
30	Toronto	- 104.30	33.10	37.09
31	Thunder Bay	17.91	75.70	43.78
32	Regina	43.48	65.96	45.27
33	Saskatoon	- 6.25	69.83	32.55
34	Calgary	88.99	45.77	55.06
35	Edmonton	24.73	43.89	39.42
36	Vancouver	- 167.89	36.55	40.53
37	Constant	133.07	43.42	93.61

See footnotes to Text Table II.

TEXT TABLE VI. Regression Coefficients and Estimated Standard Errors Dependent Variable: Clothing Expenditure

TABLEAU EXPLICATIF VI. Coefficients de régression et erreurs-types estimées variable dépendante: dépenses en vêtements

Standard errors Erreurs-types	Ratio jackknife to weighted	Sample(2) size		
Jackknife(1) repeated replicaton	Ratio jackknife/ M.C. pondérés	Taille de l'échan-tillon(2)		
Subdivision répétée jackknife(1)				N°
0.00610	5.8		Total, dépenses	1
27.77	1.2	1,408	Un adulte, sans enfants	2
50.10	1.5	396	Trois adultes, sans enfants	3
68.68	1.6	198	Quatre adultes, sans enfants	4
51.95	1.1	236	Un adulte, un enfant ou plus	5
34.34	1.2	683	Deux adultes, un enfant	6
26.63	0.9	732	Deux adultes, deux enfants	7
47.03	1.2	303	Deux adultes, trois enfants	8
70.35	1.1	98	Deux adultes, quatre enfants	9
298.62	3.1	48	Deux adultes, cinq enfants ou plus	10
74.36	1.7	205	Trois adultes, un enfant	11
72.14	1.7	207	Trois adultes, deux enfants ou plus	12
77.14	2.2	376	Autres types de famille	13
57.60	1.5	548	Âge du chef: moins de 25 ans	14
36.17	1.4	1,569	Âge du chef: 25-34 ans	15
40.37	1.6	1,234	Âge du chef: 35-44 ans	16
41.88	1.6	930	Âge du chef: 55-64 ans	17
47.48	1.6	1,230	Âge du chef: 65 ans ou plus	18
45.51	1.9	1,809	Propriétaires avec hypothèque	19
36.65	1.6	2,954	Locataires: ordinaires	20
44.23	1.0	188	Locataires: chambreurs	21
73.35	1.0	84	Locataires: sans loyer	22
58.65	1.4	261	Mode d'occupation mixte	23
34.62	0.4	355	St. John's(3)	24
33.14	0.6	371	Halifax	25
39.51	0.5	269	Saint John	26
47.54	1.0	469	Québec	27
46.37	1.4	1,000	Montréal	28
36.39	0.8	350	Ottawa	29
36.70	1.2	813	Toronto	30
43.72	0.6	351	Thunder Bay	31
46.87	0.7	341	Regina	32
32.42	0.5	353	Saskatoon	33
54.98	1.2	365	Calgary	34
39.52	0.9	518	Edmonton	35
40.20	1.1	620	Vancouver	36
92.19	2.1		Constante	37

Voir les renvois du tableau explicatif II.

dummy variable set must be dropped from the model. Accordingly, two adult, no children family type, home-owner without mortgage, age of head 45-54 years and the city of Winnipeg have been dropped. This particular set of omissions was chosen as there are numerous observations in each group so that some stability in the estimate of the constant could be anticipated. The analysis provides an opportunity to assess the use of balanced repeated replication and jackknife repeated replication in estimating standard errors of regression coefficients of both continuous and dummy variables.

It will be noted from the Text Tables II to IV that the standard error estimates produced by the balanced repeated replication and jackknife methods are generally larger than those produced by weighted least squares. This is particularly true in the case of the continuous variable: total expenditure. This reflects the author's experience gained from more simple models constructed to test the procedure.

In the case of the dummy variables the situation is not quite as straightforward. The third and fourth columns generally have values of 50% greater than column two, but this is not universally true. The dummy variables for smaller cities often have standard error estimates which are lower than those of weighted least squares. With the larger cities, the reverse is true. Why this happened is not clear, but it is possible that the result is connected with the method of assigning paired areas. With greater numbers of paired areas in a city, an increase in the ratio of the balanced repeated replication estimate to the weighted least squares estimate is experienced.

Comparisons of balanced repeated replication and jackknife methods reveal generally similar results with usually slightly smaller estimates for the latter method. This seems to be true even with the dummy variables for city despite their correlation with the method of assigning paired areas and provides some confidence in the results. Furthermore, there is no obvious relationship between the number of observations in a city and the standard error associated with that city, which seems to be in direct contrast to the situation of the weighted least squares estimates. This observation does not, however, hold up very well with the other sets of dummy variables.

From examination of the coefficients from the jackknife regression replicates, it has been observed that the estimates for the dummy variables for cities do not vary a great deal except when the replicates inside the

éliminer une variable de chaque ensemble de variables auxiliaires du modèle. C'est pourquoi on a éliminé les caractères type de famille: deux adultes - sans enfants, propriétaire - pas d'hypothèque, âge du chef: 45-54 ans et ville de Winnipeg. On a retenu cet ensemble d'omissions car il y a de nombreuses observations dans chaque groupe, de sorte que l'on peut s'attendre à une certaine stabilité de l'estimation de la constante. L'analyse permet d'évaluer les deux techniques pour l'estimation des erreurs-types des coefficients de régression des variables continues et auxiliaires.

On remarquera, à partir des tableaux explicatifs, II à IV, que les estimations de l'erreur-type obtenues par les deux méthodes sont de façon générale supérieures à celles obtenues par la méthode des moindres carrés pondérés. Cela est particulièrement évident dans le cas de la variable continue: total des dépenses. Cette différence traduit l'expérience que l'auteur a acquise en construisant des modèles plus simples pour tester la méthode.

La situation n'est pas aussi claire dans le cas des variables auxiliaires. Les troisième et quatrième colonnes ont en général des valeurs supérieures de 50 % à celles de la deuxième colonne, mais pas toujours. La variable auxiliaire des villes plus petite a souvent des estimations de l'erreur-type inférieures à celles obtenues par les moindres carrés pondérés. C'est l'inverse que l'on observe dans le cas des villes plus grandes. La raison de cette situation n'est pas claire, mais il est possible qu'il y ait là un rapport avec la façon d'affecter les régions couplées. Lorsque le nombre de régions couplées dans une ville augmente, le ratio de l'estimation obtenue par la méthode de la subdivision répétée équilibrée à celle de la technique du moindres carrés pondérés augmente également.

Une comparaison des deux méthodes fait ressortir en général des résultats semblables, la technique du jackknife donnant habituellement des estimations légèrement plus petites. Il semble que cela soit vrai aussi pour les variables auxiliaires pour les villes, en dépit de leur corrélation avec la façon d'affecter les régions couplées, ce qui permet d'avoir une certaine confiance dans les résultats. De plus, il n'y a pas de relation évidente entre le nombre d'observations dans une ville et l'erreur-type qui y est associée, ce qui semble différer totalement de la situation que l'on observe dans le cas des estimations des moindres carrés pondérés. Cette observation cependant ne se vérifie pas très bien pour ce qui est des autres ensembles de variables auxiliaires.

L'examen des coefficients obtenus pour les subdivisions utilisant la méthode de régression jackknife montrent que les estimations des variables auxiliaires des villes ne varient pas beaucoup sauf lorsque l'on prend les subdivisions

particular city are considered. This would suggest, perhaps, low correlation between the coefficient estimates of the dummy variable set and results in confidence in the stability of the model.

The similarity between the results of jackknife repeated replication and balanced repeated replication is consistent with the results published by Frankel. Further confidence in the results is yielded from this observation.

Text Tables II, III and IV also present the coefficient estimates in each case. These were estimated from the full sample $\underline{X}'\underline{X}$ and $\underline{X}'\underline{Y}$ matrices and not from the average values of the replicates. Clearly, each half sample's estimates from balanced repeated replication are unbiased as far as the coefficients are concerned. Their average consequently can be used as an estimate of the coefficients, but as the full sample estimates are not the averages of the half sample's, this is not the recommended procedure. The full sample figures are therefore the only ones provided here.

The exact distribution of the estimators of both the coefficients and the standard errors is not known at this time. This is a possible area for further research. From the point of view of hypothesis testing, however, no major problems are anticipated. Kish and Frankel(5) suggest that the estimates of regression coefficients based on clustered samples do, in fact tend towards normality, although this tendency is retarded by between cluster correlations. The "t" statistic tests used singularly on the coefficients, being quite robust will probably yield satisfactory results provided that unbiased estimators of the standard errors are available. These standard error estimates should not come from the use of traditional weighted least squares formulae.

Summary and Conclusions

In this report variance estimation techniques have been examined for use with family expenditure surveys. The methodologies for traditional Keyfitz estimators, balanced repeated replication and jackknife repeated replication were presented. Results of their use were tabulated both in the context of sample means and regression coefficients. A comparison between these results and the output of computer programs which assume a stratified simple random sample survey design was demonstrated. As was anticipated, the variances estimated do in fact vary between the Keyfitz, the balanced repeated replication

d'une ville particulière. Il semblerait qu'il s'agit là peut-être d'une faible corrélation entre les estimations des coefficients de l'ensemble de variables auxiliaires, ce qui fait que l'on peut avoir confiance en la stabilité du modèle.

La similitude entre les résultats obtenus grâce aux deux méthodes est conforme aux résultats publiés par Frankel. On peut donc avoir davantage confiance dans les résultats.

Les tableaux explicatifs II, III et IV présentent également les estimations des coefficients dans chaque cas. Ces coefficients ont été estimés à partir des matrices complètes $\underline{X}'\underline{X}$ et $\underline{X}'\underline{Y}$ de l'échantillon, et non à partir des valeurs moyennes des subdivisions. Il est clair que les estimations de chaque demi-échantillon obtenues par subdivision répétée équilibrée sont sans biais pour ce qui est des coefficients. On peut donc utiliser leur moyenne comme une estimation des coefficients, mais comme les estimations de l'échantillon complet ne sont pas les moyennes des demi-échantillons, ce n'est pas une procédure recommandée. Les chiffres de l'échantillon complet sont donc les seuls présentés ici.

La distribution exacte des estimateurs des coefficients et des erreurs-types n'est pas connue à l'heure actuelle. C'est un domaine d'étude possible. Toutefois, on ne prévoit aucun problème grave en ce qui concerne les tests d'hypothèses. Kish et Frankel(5) pensent que les estimations des coefficients de régression basés sur des échantillons en grappes tendent en fait vers la normalité, bien que cette tendance soit gênée par la corrélation entre les grappes. Les tests de la statistique "t" appliqués de façon singulière aux coefficients, étant fort robustes, vont probablement donner des résultats satisfaisants à condition que l'on dispose d'estimateurs sans biais des erreurs-types. Ces estimations de l'erreur-type ne doivent pas provenir de l'utilisation de formules rattachées aux moindres carrés pondérés traditionnels.

Résumé et conclusions

Nous venons d'examiner des techniques d'estimation de la variance destinées à être appliquées aux enquêtes sur les dépenses des familles. Nous avons présenté les méthodologies des estimateurs traditionnels de Keyfitz, de la subdivision répétée équilibrée et de la subdivision répétée du type jackknife. Les résultats de leur utilisation ont été mis en tableaux à la fois au chapitre des moyennes des échantillons et des coefficients de régression. On a effectué une comparaison entre ces résultats et les résultats de programmes informatiques utilisant un échantillon aléatoire stratifié simple. Comme on s'y attendait, les estimations de la variance ont

and jackknife repeated replication on one hand and the results obtained using formulae pertaining to stratified random samples on the other. This variation is considered sufficient to strongly advise against the use of statistics which do not consider the actual sample design.

As a practical measure, if the actual design is not considered, statistics which understate the standard errors and consequently produce inaccurate statements of confidence intervals, of coefficients of variation and of hypothesis tests can be expected. Specifically, confidence intervals which are too short and coefficients of variation which are too small will result. One can also anticipate that the power of hypothesis tests will be greatly reduced. A researcher is strongly advised to avoid this situation.

From a theoretical standpoint, a choice between jackknife and balanced repeated replication is difficult. The jackknife method is easier to use computationally but more research is needed to provide theoretical criteria for making this choice.

Additional empirical work is necessary to test the robustness of the procedures. More work with historical and future Family Expenditure Surveys is warranted. This work could include an investigation of the method of delineating the sample into paired areas. Also of interest is the examination of results coming from surveys which do not use a labour force sample design. Results consistent with those in this report would greatly increase confidence in the methods described here.

varié en fait entre les résultats des trois méthodes susmentionnées et ceux obtenus par l'emploi de formules reliées aux échantillons aléatoires stratifiés. Les écarts sont considérés suffisamment importants pour déconseiller fortement l'utilisation de statistiques qui ne prennent pas en compte le plan de sondage réel.

Comme mesure pratique, si l'on ne prend pas en compte le plan de sondage réel, on peut s'attendre à obtenir des statistiques qui sous-évaluent les erreurs-types, et par conséquent, donnent des énoncés erronés des intervalles de confiance, des coefficients de variation et des tests d'hypothèses. Plus précisément, on obtiendra des intervalles de confiance trop courts et des coefficients de variation trop faibles. On peut aussi s'attendre à ce que la puissance des tests d'hypothèses soit considérablement réduite. On recommande donc fortement au chercheur d'éviter une telle situation.

Sur le plan théorique, il est difficile de choisir entre la méthode du jackknife et celle de la subdivision répétée équilibrée. Pour ce qui est des calculs, la première est plus facile d'emploi, mais il faut faire d'autres recherches afin d'obtenir les critères théoriques d'un tel choix.

D'autres recherches empiriques sont nécessaires pour tester la robustesse des méthodes. Les enquêtes sur les dépenses passées et futures des familles vont susciter d'autres travaux. Ces travaux pourraient comprendre une étude de la façon de délimiter l'échantillon en régions couplées. L'étude des résultats d'enquêtes n'utilisant pas un plan de sondage de la population active peut également se révéler intéressante. Des résultats conformes à ceux présentés ici augmenteraient de façon considérable la confiance dans les méthodes que nous venons de décrire.

BIBLIOGRAPHY

1. Draper, N. and Smith, H., "Applied Regression Analysis", John Wiley and Sons (1966).

2. Frankel, M.R., "Inference from Survey Samples: an Empirical Investigation", Institute for Social Research, University of Michigan, Ann Arbor, Michigan (1971).

3. Keyfitz, N., "Estimates of Sampling Variance where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52 (1957), pp. 503-510.

4. Kish, L., "Survey Sampling", John Wiley and Sons (1967).

5. Kish, L., and Frankel, M.R., "Balanced Repeated Replication for Standard Errors", Journal of the American Statistical Association, Vol. 65 (1970), pp. 1071-1094.

6. McCarthy, P.J., "Pseudo-Replication: Half-Samples", International Statistics Institute Review, Vol. 33 (1969), pp. 239-264.

7. Plackett, R.L., and Burman, P.J., "The Design of Optimum Multifactorial Experiments", Biometrika 33 (1946), 305-325.

8. Statistics Canada, "Urban Family Expenditure, 1974", Publication 62-544, Occasional (1977).

9. Statistics Canada, "Methodology of the Canadian Labour Force Survey, 1976", Publication 71-526, Occasional (1977).

BIBLIOGRAPHIE

1. Draper, N. et Smith, H., "Applied Regression Analysis", John Wiley and Sons (1966).

2. Frankel, M.R., "Inference from Survey Samples: an Empirical Investigation", Institute for Social Research, University of Michigan, Ann Arbor, Michigan (1971).

3. Keyfitz, N., "Estimates of Sampling Variance where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52 (1957), pp. 503-510.

4. Kish, L., "Survey Sampling", John Wiley and Sons (1967).

5. Kish, L., et Frankel, M.R., "Balanced Repeated Replication for Standard Errors", Journal of the American Statistical Association, Vol. 65 (1970), pp. 1071-1094.

6. McCarthy, P.J., "Pseudo-Replication: Half-Samples", International Statistics Institute Review, Vol. 33 (1969), pp. 239-264.

7. Plackett, R.L., et Burman, P.J., "The Design of Optimum Multifactorial Experiments", Biometrika 33 (1946), 305-325.

8. Statistique Canada, "Dépenses des familles urbaines, 1974", publication n° 62-544, hors série (1977).

9. Statistique Canada, "Méthodologie de l'enquête sur la population active du Canada 1976", publication n° 71-526, hors série (1977).

POOR

QUALITY

ORIGINAL